|  |
| --- |
| A white square with a blue border  Description automatically generated |
|  |

|  |  |
| --- | --- |
| **Synthetic Data for Bias Mitigation (BIAS Project)** | |
| Type of work: **Term Project 2** | |
|  | |
| Degree programme: | **Master of Science in Engineering (MSE)**  **specialisation Data Science** |
| Specialisation: | **Data Science** |
| Student: | **Shilpi Garg (gargs1)** |
| Project: | **PA2** |
| Project Advisor: | **Dr. Kurpicz-Briki Mascha** |
| Co-Advisor: | **Dr. Alexandre Puttick** |
| Date: |  |

**Contents:**

1. **Abstract**
2. **Introduction**
3. **Materials and Methods**
   1. **Datasets**
   2. **Data Structure**
   3. **Experimental setup**
4. **Results**
5. **Discussion**

**6.1**

**6.2 Limitations**

**6.3 Future Work**

**6.4**

1. **Bibliography**
2. **Declaration of Authorship**
3. **Abstract:**

The Horizon Europe project BIAS explores how societal biases manifest in AI systems and large language models, particularly within the context of the labor market. As part of this initiative, this subproject focuses on bias detection and mitigation in the generation and evaluation of CVs and cover letters. The primary objective is to develop a modular synthetic data generation framework that enables the systematic creation of job application documents—including job ads, CVs, and cover letters—while carefully controlling for demographic variables, linguistic patterns, and skill profiles. This synthetic dataset is intended to serve as a benchmark for evaluating and improving fairness in downstream natural language processing (NLP) applications.

In this context, synthetic data for bias mitigation refers to the deliberate generation of artificial application materials that allow for fine-grained manipulation of attributes such as gender, ethnicity, age, and socio-economic background. This ensures that models trained or tested using such data do not perpetuate or exacerbate existing societal biases. The framework supports the injection of sensitive terms (e.g., “female,” “disabled”) and proxy terms (e.g., “parental leave,” “youth leader”) into the documents to test the model’s predictive behavior in response to these variables. These terms are often correlated with demographic markers and may trigger unintended discriminatory outcomes, making them critical for bias stress-testing and mitigation strategies.

**Keywords**- NLP, Synthetic Data, RAG, data augmentation, bias mitigation, sensitive attributes, proxy words, machine learning, LIME

1. **Introduction:**

The integration of artificial intelligence (AI) into recruitment processes has transformed how organizations evaluate job applicants. While AI offers scalable tools for resume screening and candidate evaluation, growing evidence suggests that these systems risk perpetuating or amplifying **gender biases** embedded in training data and model architectures ([Fabris et al., 2025).](https://dl.acm.org/doi/10.1145/3696457) This is particularly critical during the early stages of hiring, where automated models interpret CVs and cover letters—documents that inherently reflect individual self-presentation strategies shaped by societal norms ([Marti Marcet, 2023](https://studenttheses.uu.nl/bitstream/handle/20.500.12932/44294/Applied_Data_Science_Thesis_Sara.pdf)). AI systems trained on real-world data are vulnerable to perpetuating historical societal biases, particularly gender biases embedded in language and occupational roles. This study is part of the broader Horizon Europe BIAS project, which aims to examine and mitigate such biases in AI systems deployed within labor market contexts.

**2.1. Related work:**

**2.1.1 Gender Disparities in Training Data**

One of the foundational challenges in mitigating bias in AI hiring systems is the representational imbalance in training datasets. This issue is evident in the widely used BiasBios dataset—a corpus of over 390,000 biographies labeled with gender and occupation, contain embedded demographic skews that can influence downstream decision-making. As visualized in Figure 1 and 2, a disproportionate representation exists between male and female entries, with 53.9% identified as male (213,543) and 46.1% as female (182,646). While this may seem modest, such imbalances become more problematic when they intersect with gender-stereotyped occupational labels, where roles like “nurse” or “yoga teacher” show over **80% female representation**, while positions such as “software engineer” or “rapper” are male-dominated ([Mansouri et al., 2024](https://www.mdpi.com/2673-2688/5/1/19); [Mihaljević et al., 2022](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0274312))

Further disaggregation by profession (Figure 3) illustrates that certain occupations are strongly gender-coded in the dataset. For instance, female representation is overwhelmingly dominant in roles such as:

* Dietitian (93%)
* Nurse (91%)
* Yoga Teacher(85%)
* Interior Designer (81%)

In contrast, male-dominated professions include:

* Rapper (90%)
* Dj(86%)
* Surgeon (85%)
* Software Engineer (85%)

These imbalances are more than representational; they shape model behavior. For instance, ([Njoto et al.,2022](https://aclanthology.org/2022.nlpcss-1.15/)) illustrate that pretrained NLP models frequently learn **semantic associations** between gendered terms and specific job roles, leading to **disparate treatment of candidates** in algorithmic hiring pipelines. These disparities reinforce stereotypical gender-role associations and pose serious risks when used in training AI systems for job screening or candidate recommendation. Without intervention, models trained on such data are likely to learn and reproduce biased mappings between gendered language cues and professional competence.

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 1:** Gender distribution on whole dataset.

A graph with a blue and pink bar chart

AI-generated content may be incorrect.

**Figure 2:** Gender distribution in BiasBios dataset

A graph with different colors and text

AI-generated content may be incorrect.

**Figure 3:** Gender Representations in various Professions

**2.1.2. Synthetic Data for Bias Mitigation**

To counteract such biases, synthetic data generation has emerged as a promising solution ([Peña et al., 2023](https://link.springer.com/content/pdf/10.1007/s42979-023-01733-0.pdf)),([Frazzetto, 2025](https://tesidottorato.depositolegale.it/bitstream/20.500.14242/202137/1/tesi_Paolo_Frazzetto_fin.pdf)). This project proposes a modular synthetic data generation framework for constructing balanced and bias-aware training and evaluation datasets. Specifically, our framework focuses on the systematic generation of job application materials We generate artificial CVs and cover letters that allow controlled variation of:

* Demographic attributes (e.g., gender markers, names)
* Linguistic features
* Sensitive/proxy terms (e.g., “maternity leave”, “first-generation college graduate”): **Proxy words** are terms that the model uses as indirect signals (often unintended) to predict sensitive attributes (like gender, race, or biases).

Related detailed work and experiments are discussed in Materials and Methods section.

The synthetic data are designed to vary along critical demographic dimensions while preserving semantic and professional coherence. These terms are not inherently biased, but their correlated associations with protected characteristics make them essential for bias stress-testing. This aligns with recent approaches in the literature where **counterfactual generation** is used to test and mitigate model responses ([Kumar et al., 2023](https://ceur-ws.org/Vol-3490/RecSysHR2023-paper_7.pdf)).

The objectives of this study are threefold:

* Quantify the extent of gender bias in existing dataset (Biasbios) used in NLP hiring systems.
* Develop a reproducible pipeline for generating demographically controlled synthetic application materials.
* Evaluate the behavior of existing NLP models against this synthetic dataset using LIME classifier to assess and mitigate their bias responses.

By leveraging controlled synthetic data and fairness-driven evaluation protocols, this work contributes to the development of transparent, accountable, and equitable AI tools in recruitment pipelines. This approach is aligned with emerging principles of **Human-Centered AI** ([Bartl et al., 2025](https://dl.acm.org/doi/full/10.1145/3700438)) and contributes to a growing body of work advocating for **accountable AI in employment technologies** ([Chaturvedi & Chaturvedi, 2025](https://arxiv.org/abs/2504.21400)),([Serna et al., 2023](https://link.springer.com/content/pdf/10.1007/s42979-023-01733-0.pdf)). Moreover, it provides a scalable blueprint for testing bias sensitivity in downstream NLP applications, especially in high-stakes domains.

1. **Materials and Methods:**
   1. **BiasBios Dataset:**

BiasBios is a large-scale, human-labeled dataset developed to facilitate the study and mitigation of gender bias in machine learning systems, particularly in automated occupation classification tasks. Introduced by [De-Arteaga et al](https://arxiv.org/abs/1901.09451). in their seminal work “Mitigating Unwanted Biases with Adversarial Learning” (AIES 2019), the dataset comprises approximately 400,000 short biographies scraped from English-language publicly available professional platforms. It covers 28 diverse occupations including healthcare, tech, education, and law. It includes binary gender annotations inferred from first names or pronouns. BiasBios provides a realistic benchmark for evaluating fairness-aware algorithms and debiasing techniques. Due to the inherent occupational and gender imbalances within the dataset—mirroring real-world labor distributions—it is widely used in fairness research to analyze, detect, and mitigate societal biases perpetuated by machine learning classifiers and natural language processing (NLP) models.

* 1. **Models:**

We use large language models like **LLaMA-2-7B** and **LLaMA-3.2-3B** to automatically generate professional documents such as **CVs**, **job advertisements**, and **cover letters**. These models are fine-tuned or prompted with structured data and sample bios to produce tailored, coherent, and contextually appropriate content. By leveraging their ability to understand and generate human-like language, we can create customized outputs that reflect specific job roles, skills, and candidate profiles while also enabling scalable document generation for hiring or application pipelines.

* + 1. **Model LLaMA-2-7B-Chat (Meta AI, 2023)**

The LLaMA-2-7B-Chat model is a 7-billion parameter instruction-tuned large language model released by Meta AI as part of the LLaMA-2 family. It is specifically fine-tuned for dialogue-based interactions and chat-style completions. Built on a transformer architecture, LLaMA-2 was trained on 2 trillion tokens of publicly available data and fine-tuned using Reinforcement Learning from Human Feedback (RLHF). Its context window is 4096 tokens. It performs competitively with other open-source models like Falcon-7B and Mistral-7B on instruction-following tasks and is particularly suited for structured generation tasks such as CV/cover letter synthesis and bias analysis.

* + 1. **Model LLaMA-3.2-3B-Instruct (Meta AI, 2024)**

The LLaMA-3.2-3B-Instruct is a smaller, instruction-tuned version of the LLaMA-3 family with 3.2 billion parameters. While significantly lighter than the 7B variant, it offers excellent efficiency for low-latency inference and real-time evaluation tasks. It’s tuned on a combination of supervised datasets and synthetic prompts using chain-of-thought and system-level instruction formats. Its ideal for **real-time model probing**, fairness testing, and synthetic data generation under constrained resources. It can run on consumer GPUs or edge servers.

* 1. **Hardware specifications:**

The models were hosted and run on a **local workstation/server** with the following **GPU and CPU specifications**: **CPU**: AMD Ryzen (32 Cores, 64 Threads, 3.5 GHz base clock). High-core-count CPU is used for efficient token pre-processing and parallel sampling. **GPU**: NVIDIA A100 40GB (PCIe version). CUDA Capability: 8.0. FP16/TF32 optimization for faster matrix multiplications during transformer forward passes.

* 1. **Experimental Setups:**

Classifier is employed with a transformer-based architecture for text classification using the **BERT** (Bidirectional Encoder Representations from Transformers) model ([Jacob Devlin et al., 2018](https://arxiv.org/abs/1810.04805)). Specifically, we utilized the **bert-base-uncased** variant available through the **Hugging Face Transformers** library ([Wolf et al., 2020](https://aclanthology.org/2020.emnlp-demos.6/)), integrated via the **AutoModelForSequenceClassification** class. The input data is pre processed using the associated tokenizer with a fixed max\_length padding strategy to ensure uniform sequence length across all samples. The dataset was partitioned into training and test sets using a standard 80/20 train–test split. For training the model, we leveraged the Trainer API along with **TrainingArguments** from the Transformers framework, which allows streamlined training and evaluation workflows.

Evaluation of model performance was conducted using a suite of classification metrics from the **scikit-learn** package, including **accuracy**, **precision**, **recall**, and **F1-score**, to comprehensively assess the predictive performance. Additionally, a confusion matrix was computed to visualize the distribution of classification results, which was rendered using a heatmap generated with the seaborn visualization library.

To ensure interpretability of the model predictions, we employed LIME (**Local Interpretable Model-agnostic Explanations**) ([Ribeiro et al., 2016](https://arxiv.org/pdf/1602.04938)), a post-hoc explanation technique designed to provide insight into individual predictions made by black-box models. It explains the model’s behavior only around the single instance being analyzed, not globally. This is achieved by generating perturbed samples in the neighborhood of the instance and observing the black-box model's predictions on these samples. Perturbed samples closer to the original instance are weighted more heavily. The coefficients of this surrogate model indicate the importance of different input features (e.g., words in a sentence) for the prediction.

The project explores the job classifier relying on words correlated with gender instead of job qualifications. When explaining the model's prediction for an instance (e.g., a job ad or a CV), LIME highlights the words that contributed most to the model’s decision. By analyzing multiple instances where the model predicts a particular label (e.g., “male” or “female”), one can observe if certain words consistently have high importance. Proxy words often appear as highly weighted features in the local explanations, revealing that the model may be relying on these words as shortcuts or proxies for sensitive attributes. This identification helps in auditing and mitigating bias, for example by removing or masking proxy words or adjusting training data.

* + 1. **Gender Classifier Without Masking:**

To understand how a machine learning model makes decisions about gender classification, LIME helps us see which words in a sentence influence the model's predictions the most. We did this first **without hiding or removing any words**, so we could observe the model's behavior more naturally.

In our setup, the model predicts **gender** based on text, where **female is predicted as 1** and **male as 0**. Certain words seemed to push the model more strongly toward one gender than the other. For example, **male-related words** like *he*, *his*, *him*, *Mr.*, *Penticostal*, *aspects*, and *grant* made the model more likely to predict "male". On the other hand, **female-related words** such as *she*, *her*, *Ms.*, *Mrs.*, *Jessica*, *CPAs*, *McGinty*, and *financial* increased the chances of predicting "female". This shows that the model may be picking up on **indirect gender clues**, or **proxy words**, which are often tied to cultural roles or job titles.

|  |  |
| --- | --- |
| Top 10 Male words | Top 10 Female words |
| ('He', 3.34961242228)  ('he', 1.9166196645)  ('his', 0.9823642110)  ('him', 0.59896187373)  ('Mr', 0.212177381359)  ('For', 0.15553675710)  ('an', 0.139449532486)  ('services', 0.127056438325)  ('help', 0.12677291047)  ('Penticostal', 0.126314986395) | ('She', -20.334972885)  ('her', -6.5985962079)  ('she', -5.3634170559)  ('Her', -2.318591337)  ('Ms', -1.21109984780)  ('Mrs', -0.54779522537)  ('Jessica', -0.54219316489)  ('herself', -0.51449886288)  ('Their', -0.488303965837)  ('CPAs', -0.47497754571) |

A screenshot of a computer

AI-generated content may be incorrect.

**Figure 4:** Gender classifier using LIME (without masking) for **Figure 5:** List of top 10 Gender words without masking.

‘accountant’ Bio.

* + 1. **Gender Classifier with lowercase text and without Masking:**

The dataset text was converted to **lowercase** before being analyzed, which ensures consistency and avoids discrepancies due to capitalization.

In our setup, the model predicts gender based on text, where **female is predicted as 1** and **male as 0**. Certain words clearly influenced the model’s predictions more than others. For instance, **male-associated words** like *he*, *his*, *has*, *him*, *firm*, *professional*, *post*, *aspects*, and *department* made the model more likely to predict "male." In contrast, **female-associated words** such as *she*, *her*, *ms*, *jessica*, *husband*, *herself*, *carla*, *women*, *joining*, *financial*, and *donnelley* increased the likelihood of predicting "female."

These patterns highlight how text-based models can learn and reproduce **gender biases** based on word usage, job roles, and social cues embedded in the data.

|  |  |
| --- | --- |
| Top 10 Male words | Top 10 Female words |
| Top 10 male words:  ('he', 7.6871099670)  ('his', 2.96870740451)  ('has', 0.9409299788)  ('him', 0.53055096147)  ('in', 0.456975840817)  ('and', 0.423679987703)  ('is', 0.369853515312)  ('firm', 0.27047083722)  ('of', 0.236085533458)  ('professional', 0.197072901868) | Top 10 female words:  ('she', -21.7097535060)  ('her', -7.5950752558)  ('ms', -1.43420665474)  ('jessica', -0.54344341807)  ('wendy', -0.51713271813)  ('husband', -0.48899558316)  ('herself', -0.4540737488)  ('carla', -0.42491086950)  ('catherine', -0.41852755173)  ('women', -0.38392412151) |

A screenshot of a computer

AI-generated content may be incorrect.

**Figure6:** Gender classifier using lowercase text and without masking **Figure 7:** List of top 10 Gender words(lowercase)

1. **Results:**
2. **Discussion:**
3. **Bibliography:**

 [De-Arteaga et al](https://arxiv.org/abs/1901.09451) (2019) . Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. [arxiv.org/abs/1901.09451](https://arxiv.org/abs/1901.09451)

 Bartl, M., Mandal, A., Leavy, S., & Little, S. (2025). Gender bias in natural language processing and computer vision: A comparative survey. <https://dl.acm.org/doi/full/10.1145/3700438>

 Chaturvedi, S., & Chaturvedi, R. (2025). Who Gets the Callback? Generative AI and Gender Bias. <https://arxiv.org/abs/2504.21400>

 Fabris, A., Baranowska, N., Dennis, M. J., Graus, D. et al. (2025). Fairness and Bias in Algorithmic Hiring: A Multidisciplinary Survey. <https://dl.acm.g/doi/10.1145/3696457>

 Frazzetto, P. (2025). Leveraging Deep Learning in Human Resources: Graph Neural Networks for Candidate-Job Matching. [Thesis PDF](https://tesidottorato.depositolegale.it/bitstream/20.500.14242/202137/1/tesi_Paolo_Frazzetto_fin.pdf)

 Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Rush, A. M. (2020). *Transformers: State-of-the-Art Natural Language Processing*. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 38–45. <https://aclanthology.org/2020.emnlp-demos.6>

 Kumar, D., Greif, E., Rekabsaz, N., & Schedl, M. (2023). Identifying Words in Job Advertisements Responsible for Gender Bias via Counterfactual Learning. CEUR-WS. [PDF](https://ceur-ws.org/Vol-3490/RecSysHR2023-paper_7.pdf)

 Mansouri, T., Alameer, A., & Albaroudi, E. (2024). AI Techniques for Addressing Algorithmic Bias in Job Hiring. AI, 5(1), 19. <https://www.mdpi.com/2673-2688/5/1/19>

 Marti Marcet, S. (2023). Natural Language Processing in Resume Data: The Interplay Between Gender and Occupation on Resume Writing Style. Utrecht University. [PDF](https://studenttheses.uu.nl/bitstream/handle/20.500.12932/44294/Applied_Data_Science_Thesis_Sara.pdf)

 Njoto et al. (2022). *Professional Presentation and Projected Power: A Case Study of Implicit Gender Information in English CVs.* <https://aclanthology.org/2022.nlpcss-1.15/>

 Mihaljević, H., Müller, I., Dill, K., & Yollu-Tok, A. (2022). Towards Gender-Inclusive Job Postings: A Data-Driven Comparisonof augmented writing technologies. PLOS ONE, 17(12). <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0274312>

 Peña, A., Serna, I., Morales, A., Fierrez, J., & Ortega, A. (2023). Human-Centric Multimodal Machine Learning: Recent Advances and Testbed on AI‑Based Recruitment. SN Computer Science, 4(5). [s42979-023-01733-0.pdf](https://link.springer.com/content/pdf/10.1007/s42979-023-01733-0.pdf)

 Ribeiro, M., Singh, S., & Guestrin, C. (2016). *Why should I trust you?": Explaining the predictions of any classifier.* [1602.04938v3.pdf](https://arxiv.org/pdf/1602.04938)

 Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* <https://arxiv.org/abs/1810.04805>

1. **Declaration of Authorship**

I hereby certify that I composed this work completely unaided, and without the use of any other sources or resources other than those specified in the bibliography. All text sections not of my authorship are cited as quotations and accompanied by an exact reference to their origin.

Place, date: BFH, Biel

Signature: Shilpi Garg